On the Compositionality and Semantic Interpretation of English Noun Compounds

Corina Dima
Collaborative Research Center 833
University of Tübingen, Germany
corina.dima@uni-tuebingen.de

Abstract

In this paper we present a study covering the creation of *compositional distributional representations* for English noun compounds (e.g. *computer science*) using two compositional models proposed in the literature. The compositional representations are first evaluated based on their similarity to the corresponding corpus-learned representations and then on the task of automatic classification of semantic relations for English noun compounds. Our experiments show that compositional models are able to build meaningful representations for more than half of the test set compounds. However, using pre-trained compositional models does not lead to the expected performance gains for the semantic relation classification task. Models using compositional representations have a similar performance as a basic classification model, despite the advantage of being pre-trained on a large set of compounds.

1 Introduction

Creating word representations for multiword expressions is a challenging NLP task. The challenge comes from the fact that these constructions have "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002). A good example of such challenging multiword expressions are noun compounds (e.g. *finger nail*, *health care*), where the meaning of a compound often involves combining some aspect or aspects of the meanings of its constituents.

Over the last few years distributed word representations (Collobert et al., 2011b; Mikolov et al., 2013; Pennington et al., 2014) have proven very successful at representing single-token words, and there have been several attempts at creating compositional distributional models of meaning for multi-token expressions, in particular adjective-word phrases (Baroni and Zamparelli, 2010), determiner phrases (Dinu et al., 2013b) or verb phrases (Yin and Schütze, 2014).

Studying the semantics of multiword units, and in particular the semantic interpretation of noun compounds has been an active area of research in both theoretical and computational linguistics. Here, one train of research has focused on understanding the mechanism of compounding by providing a label for the relation between the constituents (e.g. in *finger nail*, the *nail* is PART OF the *finger*) as in (Ó Séaghdha, 2008; Tratz and Hovy, 2010) or by identifying the preposition in the preferred paraphrase of the compound (e.g. *nail **of** the finger*) as in (Lauer, 1995).

In this paper, we explore compositional distributional models for English noun compounds, and analyze the impact of such models on the task of predicting the compound-internal semantic relation given a labeled dataset of compounds. At the same time, we analyze the results of the compositional process through the lens of the semantic relation annotation, in an attempt to uncover compounding patterns that are particularly challenging for the compositional distributional models.

2 Context and Compound Interpretation

There are two possible settings for compound interpretation: *out-of-context interpretation* and *context-dependent interpretation*.

Bauer (1983, pp. 45) describes a continuum of types of complex words, arranged with respect to their formation status and to how dependent their interpretation is on the context: (i)"*nonce formations*, coined by a speaker/writer on the spur of the moment to cover some immediate need", where there is a large ambiguity with respect to the meaning of the compound which cannot be resolved without the immediate context (e.g. Nakov's (2013) example compound *plate length*,

Proceedings of the 1st Workshop on Representation Learning for NLP, pages 27–39,
Berlin, Germany, August 11th, 2016.

for which a possible interpretation in a given context could be *what your hair is when it drags in your food*); (ii) *institutionalized lexemes*, whose potential ambiguity is canceled by the frequency of use and familiarity with the term, and whose more established meaning could be inferred based on the meanings of the constituents and prior world experience, without the need for an immediate context (e.g. *orange juice*); (iii) *lexicalized lexemes*, where the meaning has become a convention which cannot be inferred from the constituents alone and can only be successfully interpreted if the term is familiar or if the context provides enough clues (e.g. *couch potato*[1]).

The available datasets we use (described in Section 3) are very likely to contain some very low frequency items of type (i), whose actual interpretation would necessitate taking the immediate context into account, as well some highly lexicalized compounds of type (iii), where the meaning can only be deduced from context. Nevertheless, because of a lack of annotated resources that provide the semantic interpretation of a compound together with its context, we will focus on the out-of-context interpretation of compounds.

3 Datasets

3.1 English Compound Dataset for Compositionality

The English compound dataset used for the composition tests was constructed from two existing compound datasets and a selection of the nominal compounds in the WordNet database. The first existing compound dataset was described in (Tratz, 2011) and contains 19158 compounds[2]. The second existing compound dataset was proposed in (Ó Séaghdha, 2008) and contains 1443 compounds[3].

Additional compounds were collected from the WordNet 3.1 database files [4], more specifically from the noun database file `data.noun`. The WordNet compound collection process involved 3 steps: (i) collecting all candidate compounds, i.e. words that contained an underscore or a dash (e.g. *abstract_entity*, *self-service*); (ii) filtering out candidates that included numbers or dots, or had more than 2 constituents; (iii) filtering out candidates where either one of the constituents had a part-of-speech tag that was different from `noun` or `verb`. The part-of-speech tagging of the candidate compounds was performed using the *spaCy* Python library for advanced natural language processing[5]. The reason for allowing both `noun` and `verb` as accepted part-of-speech tags was that given the extremely limited context available when PoS-tagging a compound the tagger would frequently label as `verb` multi-sense words that were actually nouns in the given context (e.g. *eye drop*, where *drop* was tagged as a verb). The final compound list extracted from WordNet 3.1 contained 18775 compounds.

The compounds collected from all three resources were combined into one list. The list was deduplicated and filtered for capitalized compounds (the Tratz (2011) dataset contained a small amount of person names and titles). A final filtering step removed all the compounds where either of the two constituents or the compound itself did not have a minimum frequency of 100 in the *support corpus* (presented later, in Section 4.1). The frequency filtering step was motivated by the assumption that the compositional process can be better modeled using "well-learned" word vectors that are based on a minimum number of contexts.

The final dataset contains 27220 compounds, formed through the combination of 5335 modifiers and 4761 heads. The set of unique modifiers and heads contains 7646 words, with 2450 words appearing both as modifiers and as heads. The dictionary for the final dataset contains therefore 34866 unique words. The dataset was partitioned into `train`, `test` and `dev` splits containing 19054, 5444 and 2722 compounds respectively.

3.2 English Compound Datasets for Semantic Interpretation

The Tratz (2011) dataset and the Ó Séaghdha (2008) dataset are both annotated with semantic relations between the compound constituents. The Tratz (2011) dataset has 37 semantic relations and 19158 compounds. The Ó Séaghdha (2008) dataset has 1443 compounds annotated with 6 coarse relation labels (ABOUT, ACTOR, BE, HAVE,

[1] a *couch potato* is not a potato, but a person who exercises little and spends most of the time in front of a TV.

[2] The dataset is part of the semantically-enriched parser described in (Tratz, 2011) which can be obtained from `http://www.isi.edu/publications/licensed-sw/fanseparser/`

[3] Available at `http://www.cl.cam.ac.uk/~do242/Resources/1443_Compounds.tar.gz`

[4] Available at `http://wordnetcode.princeton.edu/wn3.1.dict.tar.gz`

[5] `https://spacy.io/`

IN, INST). Appendix A lists the relations in the two datasets together with some example annotated compounds.

For both datasets a small fraction of the constituents had to be recoded to the artificial underscore-based form described in Section 4.1, in order to maximize the coverage of the word representations for the constituents (e.g. *database* was recoded to *data_base*).

4 Composition Models for English Nominal Compounds

A common view of natural language regards it as being inherently compositional. Words are combined to obtain phrases, which in turn combine to create sentences. The composition continues to the paragraph, section and document levels. It is this defining trait of human language, its *compositionality*, that allows us to produce and to understand the potentially infinite number of utterances in a human language.

Gottlob Frege (1848-1925) is credited with phrasing this intuition into the form of a principle, known as the Principle of Compositionality: "The meaning of the whole is a function of the meaning of the parts and their mode of combination" (Dowty et al., 1981, p.8).

The adoption of distributional vectors as a proxy for the meaning of individual words (in other words, having a "meaning of the parts") encouraged researchers to focus their attention on finding *composition models* which could act as the "mode of combination".

When applied to vector space models of language, the idea of looking for a "mode of combination" translates to finding a composition function f which takes as input some n-dimensional distributional representations for the two constituents constructed using a support corpus, $u^{corpus}, v^{corpus} \in \mathbb{R}^n$ and outputs another n-dimensional representation for the compound $p^{composed} \in \mathbb{R}^n$,

$$p^{composed} = f(u^{corpus}, v^{corpus})$$

Additionally, we consider $p^{corpus} \in \mathbb{R}^n$, the learned representation for the compound, to be the "gold standard" for the composed representation of the compound $p^{composed}$. Therefore the composition function f should minimize J_{MSE}, the mean squared error between the composed and the corpus-induced representations:

$$J_{MSE} = \sum_{i=1}^{nc} \frac{1}{n} \sum_{j=1}^{n} (p_{ij}^{composed} - p_{ij}^{corpus})^2$$

where nc is the number of compounds in our dataset.

Previous studies like (Guevara, 2010; Baroni and Zamparelli, 2010) evaluate their proposed composition functions on training data created using the following procedure: first, they gather a set of word pairs to model. Then, a large corpus is used to construct distributional representations both for the word pairs as well as for the individual words in each pair. In order to derive word pair representations the corpus is first pre-processed such that all the occurrences of the word pairs of interest are linked with the underscore character '_'. This tricks the tokenizer into considering each pair a singe-unit word, thus making it possible to record its co-occurrence statistics using the same distributional methods one would use for a genuine single-unit word.

The same methodology is applied here for creating a training dataset for compositional models using the list of compounds described in Section 3.1. The process is detailed in Section 4.1.

Next, we selected two composition functions (we also refer to them as composition models) from the ones presented in the literature:

- the *full additive* model, introduced in Zanzotto et al. (2010) (in their work this model is called the *estimated additive model*) and popularized as part of the DISSECT toolkit (Dinu et al., 2013a; Dinu et al., 2013b). In this model the two constituent vectors u and $v \in \mathbb{R}^n$ are composed by multiplying them via two square matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$. $\mathbf{A}$ and $\mathbf{B}$ are the same for every u and v, so during training we only have to estimate the parameters in two $n \times n$ matrices, making the model constant in the number of parameters. The mathematical formulation of the full additive model is presented in Eq. 1.

$$p = \mathbf{A}u + \mathbf{B}v \quad (1)$$

- the *matrix* model, introduced in Socher et al. (2011). It is a non-linear composition model where the constituent vectors $u, v \in \mathbb{R}^n$ are first concatenated, resulting in a vector

$[u; v] \in \mathbb{R}^{2n}$ and then multiplied with a matrix $\mathbf{W} \in \mathbb{R}^{2n \times n}$. The result of the multiplication is an n-dimensional vector which is passed as a final step through a non-linear function g (in this case the element-wise hyperbolic tangent `tanh`). The parameter matrix $\mathbf{W}$ which has to be estimated during the training process is the same for all the possible input vectors u and v. Since this composition function is implemented via a neural network, a bias term $b \in \mathbb{R}^n$ is added after the multiplication of the matrix $\mathbf{W}$ with the concatenated vector $[u; v]$. The complete form of this composition function is given in Eq. 2.

$$p = g(\mathbf{W}[u; v] + b) \tag{2}$$

The preference for these particular composition models is justified by their constant number of parameters with respect to the vocabulary size. This allows us to use this composition model for a significantly larger number of constituents than the one in the list of compounds it was trained on. In particular, this allows us to predict a composition vector even for the compounds that were not attested in the corpus, if their constituents are frequent enough to be part of our full vocabulary.

Both models were reimplemented using the Torch7 library (Collobert et al., 2011a), whose *nngraph* module allows for an easy creation of architectures with multiple inputs and outputs. Reimplementing the composition models is also justified by the use of trained composition models as a form of *pre-training* for the semantic interpretation models described in Section 5.

4.1 Compound-aware Word Representations

The *support corpus* for creating English word representations for compositionality experiments (referred to in Section 3.1) was obtained by concatenating the raw text from the ENCOW14AX corpus (Schäfer, 2015) and the pre-processed 2014 English Wikipedia dump described and made available in Müller and Schütze (2015). A preprocessing step similar to the one described in Müller and Schütze (2015) was applied to the concatenated corpus: the text was lowercased and the digits were replaced with 0s. An additional preprocessing step was necessary for creating compound representations. A list of compounds (described in Section 3.1) was used to recode the initial corpus such that the two-part compounds in the list would be considered a single token. The recoding process involved replacing different spelling variants of a compound - written as two separate words, contiguously or with a dash (as in *dress code*, *dresscode* or *dress-code*), as well as their respective plural forms (*dress codes*, *dresscodes*, *dress-codes*) with an artificial underscore-based form (e.g. *dress_code*). We did not, however, modify the plural first constituents (i.e. *savings account*), nor did we normalize the spelling variation which is the result of different spelling standards as in *color scheme* (American English) and *colour scheme* (British English). The result was a 9 billion words raw-text corpus with a corresponding vocabulary containing 424,014 words (both simplex words and compounds) with minimum frequency 100 (the full vocabulary had 16M words).

The raw-text corpus was the basis for training 300 dimensional word representations using the GloVe package (Pennington et al., 2014). The GloVe model was trained for 15 iterations using a 10-word symmetric context (20 words in total) for constructing the co-occurence matrix. The vector spaces were normalized to the *L2*-norm, first across features and then across samples using *scikit-learn* (Pedregosa et al., 2011).

4.2 Evaluation and Results

The parameters of the two composition models described in Section 4 were estimated with the help of the list of compounds in the `train` set described in Section 3.1 and the word representations presented in Section 4.1. We evaluated the performance of the composition models on the `test` split of the dataset, using the rank evaluation proposed by Baroni and Zamparelli (2010). Using a trained model, we generate *composed* representations for all the compounds in the `test` set. The composed representation of each compound is ranked with respect to all the 34866 unique words in the dictionary (the set of all compounds and their respective constituents) using the cosine similarity. The best possible result is when the corpus-learned representation is the nearest neighbor of the composed representation, and corresponds to assigning the rank 1 to the composed vector. Rank 2 is assigned when the corpus-learned representation is the second nearest neighbor, and so on. The cut-off rank 1000 is assigned to all the representations with a rank

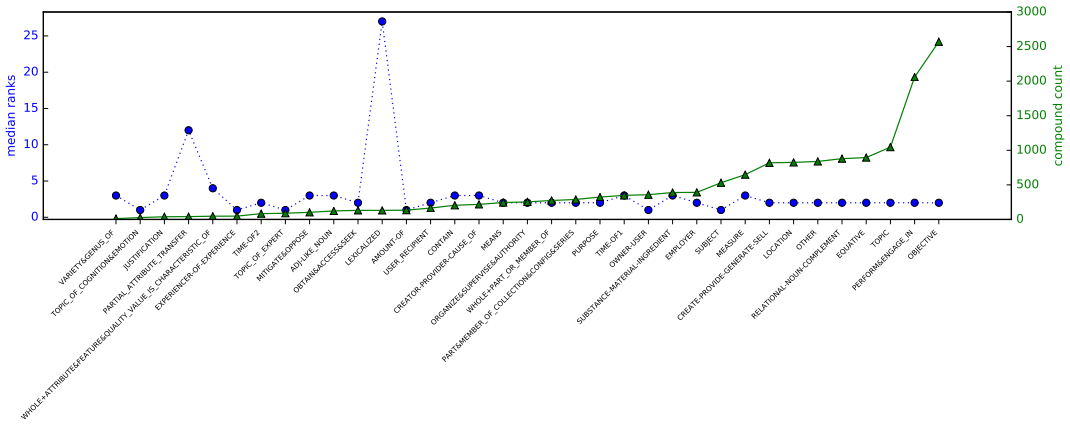

Figure 1: Semantic relations in the Tratz (2011) dataset: number of compounds labeled with a relation (green triangle) vs. the median rank assigned to their composed representations by the *full additive* model (blue circle).

$\geq$ 1000. The first, second and third quartiles (Q1, Q2/median, Q3) are then computed for the sorted list of ranks of the composed representations of the test set compounds. The result of our evaluation are displayed in Table 1.

Model	**Ranks `dev`**				**Ranks `test`**			
	Q1	Q2	Q3	Max	Q1	Q2	Q3	Max
matrix	2	5	28	1K	1	5	25	1K
full additive	1	5	28	1K	1	5	25	1K

Table 1: Composition models results: quartiles for the ranks assigned to the `dev` and `test` composed representations (lower is better).

Both composition models obtain good results on the `test` dataset with respect to the Q1, Q2, Q3 quartiles. Ranks in the 1-5 range, which were assigned to half of the `test` set compounds correspond to a well-built compound representation which resides in the expected vectorial neighborhood. For the next quarter of the data, the rank in the 6-25 range points to a representation that might still be considered reasonable depending on the application. For the last segment of ranked compounds the constructed representations are most likely incorrect. As detailed in the next paragraph, such high ranks usually suggest a difficulty in creating a compound representation based on the constituent representations and indicate that the compound belongs to a special class (e.g. lexicalized, multi-sense etc). For both models the maximum assigned rank is the cut-off rank 1000.

To put these results into perspective, the results of compositional models were interpreted through the lens of annotated semantic relations in publicly available datasets. Figure 1 plots the median rank assigned to the compounds with a particular semantic relation against the number of compounds labeled with that semantic relation in the subset of the Tratz (2011) dataset included in the compositionality dataset described in Section 3.1. The figure confirms the intuition that recovering the meaning of lexicalized compounds like *eye candy* and *basket case* is very difficult given only the constituents: the LEXICALIZED relation, which labels 131 compounds, has the median rank 27. Another difficult semantic relation for the composition model is PARTIAL_ATTRIBUTE_TRANSFER, which labels compounds such as *hairline crack* and *bullet train*, which has a median rank of 12 for its 41 compounds. The high median rank suggests that this type of attributive relation is difficult to model using distributional representations of the individual constituents, as it is based on a common attribute which is not present in the surface form of the compound (the *width* for the *hairline* and the *crack*; the *speed* for the *bullet* and the *train*).

5 Automatic Semantic Relation Classification for English Nominal Compounds

The goal of the current section is to asses the impact of composition models on the task of *automatic semantic relation classification for English nominal compounds*. The semantic relation classification task deals with predicting the correct label for the relation between the constituents of a com-

pound, given a fixed set of possible labels (e.g. the label of the relation linking *iron* to *fence* in *iron fence* is MATERIAL). The two datasets described in Section 3.2 are used as a testbed for the comparison of the composition models described in Section 4. The state of the art results for these datasets are 65.4% 5-fold cross-validation (CV) accuracy for the Ó Séaghdha dataset, obtained in Ó Séaghdha and Copestake (2013), 79.3% 10-fold CV accuracy for an unpublished version of the Tratz dataset, with 17509 noun pairs annotated with 43 semantic relations (Tratz and Hovy, 2010) and 77.70% 10-fold CV accuracy on a subset of the Tratz (2011) dataset obtained in (Dima and Hinrichs, 2015).

Our MLP architecture for semantic classification consists of two modules: the *composition module* which constructs the compound representation from the representations of its constituents and the *classification module* which takes as input the constructed compound representation and uses it as a basis for classifying the compound with respect to the semantic relations defined by each dataset.[6]

In the experiments described next the architecture of the composition module varies according to the method used for creating compound representations, while the classification module always follows the same architecture: a linear layer $W_{rel} \in \mathbb{R}^{n_c \times k}$ where n_c is the dimensionality of the compound representation and k is the number of semantic relations in the dataset, the nonlinearity `tanh` and a softmax layer that selects the "winning" semantic relation from the k possible relations. Another constant addition to the full architecture is a 0.1 dropout layer for regularization and a `reLU` nonlinearity between the composition and the classification modules.

All the described models are trained using a negative log-likelihood criterion, optimized with mini-batch Adagrad (Duchi et al., 2011) with a fixed initial learning rate (0.1, Tratz dataset; $5e{-}2$, Ó Séaghdha dataset), learning rate decay $1e-5$, weight decay $1e-5$ and a batch size of 100 as hyperparameters for the optimization process. The models are trained using early stopping with a patience of 100 epochs.

Our working hypothesis is that learning first how to compose, and then doing the semantic relation classification task should yield better results than when the composition is learned based only on the signal provided by the classification task. We expect that pre-training the composition module would make the semantic relation classification task easier and that having a good compound representation would aid its semantic interpretation.

We define as a *basic* composition module a simple architecture that takes as input u and v, the two n-dimensional constituent representations, concatenates them, and multiplies the concatenated $2n$-dimensional vector with a matrix W. Depending on the output dimensions of the model we want to compare it to, the dimensions of W will range from $\in \mathbb{R}^{2n \times n}$ to $\in \mathbb{R}^{2n \times 4n}$.

Table 2 presents the results of the classification models, grouped according to the number of parameters in the composition module. We used the *matrix* and *full additive* composition models evaluated in Section 4.2 as pre-trained composition modules.

The first two rows in Table 2 present the results of doing semantic relation classification using the *composed* compound representations as the only input to the classifier. In these settings, which are labeled *compoM*$_{300\times600}$ and *compoFA*$_{300\times600}$, the input is the composed representation as computed by the pretrained *matrix* and *full additive* composition models. The composed representations are kept fixed during the classification process. This configuration obtained the weakest results from all the tested configurations. An explanation for this result might be that the composition models perform well for only half of our test compounds, meaning that a good portion of the compounds have a potentially suboptimal representation.

In the next two rows the pre-trained composition models are *fine-tuned* for the semantic classification task (models labeled *pretrain_matrix*$_{600\times300}$ and *pretrain_fullAdditive*$_{600\times300}$). The input in this case are the initial corpus-based vectors of the two constituents.

Contrary to our hypothesis, the classification results of the *basic*$_{600\times300}$ model (the last model in the first subsection) are on par or slightly better than the previous results, where the classification used the direct or fine-tuned output of a pre-trained composition module.

This effect extends to the other settings that

[6]The code for composing representations and for doing automatic classification of semantic relations is available at `https://github.com/corinadima/gWordcomp`

Composition module	Pre-trained?	Fine-tuned?	Tratz CV	Ó Séaghdha CV
$compoM_{300\times600}$	yes	no	74.22%	57.52%
$compoFA_{300\times600}$	yes	no	73.70%	56.62%
$pretrain\_matrix_{600\times300}$	yes	yes	78.05%	59.18%
$pretrain\_fullAdditive_{600\times300}$	yes	yes	77.89%	59.18%
$basic_{600\times300}$	no	no	78.57%	59.25%
$pretrain\_matrix\_fullAdd_{600\times600}$	yes	yes	78.92%	59.39%
$basic_{600\times600}$	no	no	78.88%	59.60%
$c1c2\_compoM_{900\times900}$	yes	no	79.06%	**61.12%**
$c1c2\_compoFA_{900\times900}$	yes	no	79.07%	59.60%
$basic_{600\times1200}$	no	no	79.03%	59.60%
$c1c2\_compoMcompoFA_{1200\times1200}$	yes	no	79.16%	59.18%
$basic_{600\times2400}$	no	no	**79.36%**	58.49%

Table 2: Semantic relation classification results on the Tratz and Ó Séaghdha datasets using accuracy as a classification measure. Results obtained through 10-fold cross-validation on the Tratz dataset and 5-fold CV on the Ó Séaghdha dataset (with the original folds).

were investigated, where:

- both pre-trained composition models are used for the composition module; the compound representation is the concatenation of the two composed representations ($pretrain\_matrix\_fullAdd_{600\times600}$); even if the combined classifier outperforms each of the classifiers based on only one composition model, its results are still on par with the ones of the basic classifier with a similar number of parameters ($basic_{600\times600}$, see results in Table 2, subsection 2).

- the initial vector representations of the constituents as well as their composed representation are used as an input ($c1c2\_compoM_{900\times900}$, $c1c2\_compoFA_{900\times900}$); the composition is in this case not fine-tuned; the results on the Tratz (2011) dataset are again similar to the comparable basic model ($basic_{600\times1200}$). The $c1c2\_compoM_{900\times900}$ obtains the best overall result, 61.12%, on the Ó Séaghdha (2008) dataset.

- the input consists of the initial vector representations and both composed representations ($c1c2\_compoMcompoFA_{1200\times1200}$); the composed vectors are fixed; the results are compared to the $basic_{600\times2400}$ model (again, with a similar number of parameters). This last section contains the best overall result for the Tratz (2011) dataset, 79.36%, obtained by the $basic_{600\times2400}$ model.

To understand this unexpected result we analyzed the predictions made by the best performing classification models, $basic_{600\times2400}$ and $c1c2\_compoMcompoFA_{1200\times1200}$, on the Tratz (2011) dataset. The analysis targeted the distribution of errors per semantic relation for each of the two classifiers. As the distribution of compounds labeled with a particular semantic relation is rather skewed, we found it more informative to look at the percentage of errors for each class (shown in Figure 2) rather than at the absolute error values.

A first conclusion that can be drawn from this figure is that the two models have roughly the same distribution of errors: both struggle the most with the semantic relations with a low compound count (left side of the figure) and with the class of lexicalized compounds. In addition, even some of the relations with more than 500 labeled examples (starting from SUBSTANCE-MATERIAL-INGREDIENT) remain difficult to identify (in particular the heterogeneous OTHER relation, which labels compounds whose relation is not covered by the rest of the inventory, and the EQUATIVE relation, which labels compounds based on subtype or logical-and relations, i.e. *mozzarella cheese*, *female owner*).

An analysis of the classification errors revealed that both classifiers actually struggle to generalize above the lexical level. If a word has the majority

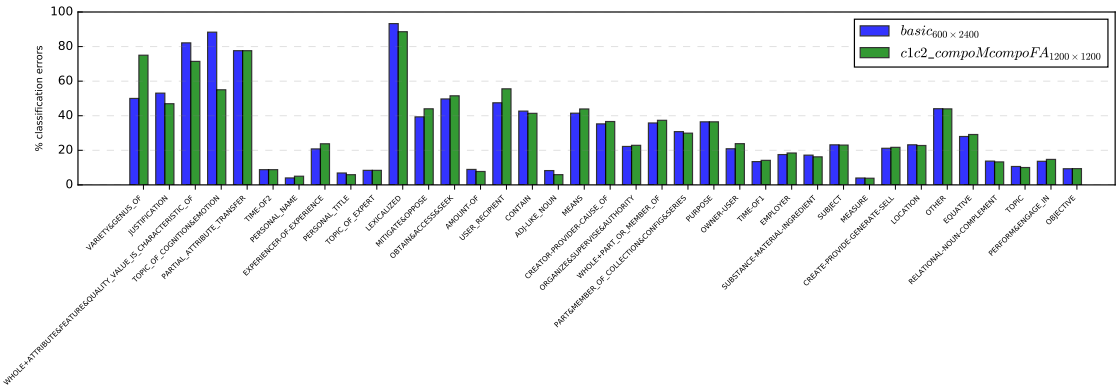

Figure 2: Error analysis for semantic relation classification on the Tratz (2011) dataset: the percentage of errors for each semantic class for the $basic_{600\times2400}$ (blue, left) and the $c1c2\_compoMcompoFA_{1200\times1200}$ (green, right) models. The semantic relations are sorted by compound count (low count to the left, high count to the right).

of compounds labeled with a relation (e.g. TOPIC for compounds with *guide*: *travel guide*, *fishing guide*), other compounds with the head *guide* will be assigned the same relation (e.g. *user guide* is labeled TOPIC although the correct relation is USER_RECIPIENT). This phenomenon where the classifier memorizes lexical associations between words in particular slots and classification labels as opposed to learning relations between the words in the two slots is referred to in Levy et al. (2015) as *lexical memorization*. To get a sense of how this phenomenon affects our classification task we plot in Figure 3 two ratios for every semantic relation in the Tratz (2011) dataset: the number of *distinct modifiers* over the total number of compounds and the number of *distinct heads* over the total number of compounds. A small ratio indicates that a large subset of the compounds labeled with a particular semantic relation share a common constituent: for example, the ADJ-LIKE_NOUN subclass has only 7 distinct modifiers for 254 compounds, resulting in a very low modifier ratio (0.03). Similarly the AMOUNT_OF subclass has 168 compounds with 15 heads (head ratio: 0.09).

Comparing Figure 3 to Figure 2, one can observe that the majority of the classes with either a low head ratio or a low modifier ratio also have the lowest percentage of errors per class. This is the case for relations like TIME_OF2, TOPIC_OF_EXPERT, AMOUNT-OF, ADJ-LIKE_NOUN and MEASURE, all of which have under 10% error rate. A notable exception is the PERSONAL_NAME semantic relation for which the classifiers manage to have a small error rate even with very diverse modifiers and heads (both modifier and head ratio is 0.96). A more realistic estimate of the actual performance of the classifiers are the semantic relations which have both a larger number of compounds and a more diverse set of constituents, like in the case of USER_RECIPIENT, CREATOR_PROVIDER_CAUSE-OF, WHOLE+PART_OR_MEMBER_OF or PURPOSE, which have a 40-60% error rate.

As a concluding point, the best results in our study are comparable to the respective state-of-the-art counterparts (79.3%/77.70% accuracy vs. 79.36% accuracy on the Tratz data; 65.4% vs. 61.12% on the Ó Séaghdha data). However, it must be taken into account that in this study the only available information for the classifiers comes from the word embeddings themselves, and from the correlations learned in the composition process. By contrast the classifiers used in (Tratz and Hovy, 2010; Tratz, 2011) relied on an extensive feature set which included information from the WordNet (hypernyms, synonyms, gloss, part-of-speech indicators; "lexicalized" indicator if the compound had an WN entry as a single term), Roget's thesaurus, surface-level features and n-gram features extracted from the Web 1T corpus. The state-of-the-art of the Ó Séaghdha (2008) dataset is based on both lexical features (for the individual constituents, constructed on the basis of dependency relations) and relational features (for the typical interactions of constituents, constructed on the basis of contexts where the constituents appear together as separate words). The distributional representations we use as input are likely to cap-

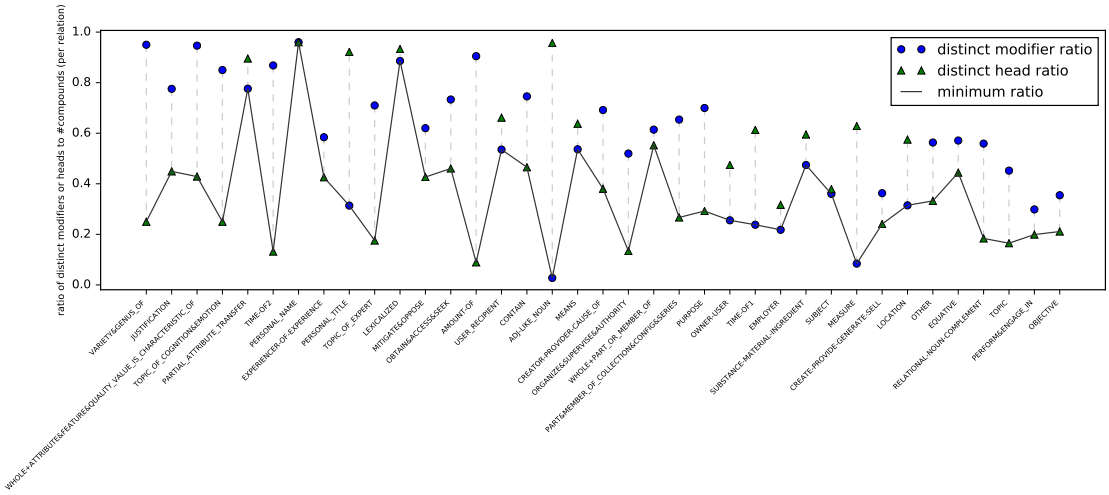

Figure 3: Diversity of modifiers and heads per relation: a low ratio for either the modifier (blue circle) or the head (green triangle) correlates with a small error rate for the classification task.

ture both lexical and relational aspects, but do not explicitly model pairwise constituent interactions.

6 Conclusions

In this paper we have presented a study covering the creation of compositional distributional representations for English noun compounds. The representations created by the compositional models were further evaluated on the task of automatic semantic relation classification for English noun compounds, using two preexisting annotated datasets. The experiments are, to the best of our knowledge, the first compositional investigations focusing on English noun compounds. The composition models have a good performance and manage to build meaningful composed vectors for half of the test set compounds.

The investigation of semantically annotated compound datasets revealed that composition models cannot represent compounds with lexicalized meaning. This suggests that the representations of compounds where the meaning of the whole is substantially different from the one of the parts should be learned directly from corpus co-occurence data. Another vocabulary-related observation concerns the extensive pre-processing necessary to create distributional representations for compounds. Spelling variation (e.g. *health care*, *health-care*, *healthcare*) artificially creates separate forms with the same meaning. Such forms should be identified and collapsed back to a single meaning representation when creating vector space models of language.

The semantic relation classification experiments showed that state-of-the-art composition models must be further refined before they can be of use for downstream semantic tasks. In our experiments compositional models were unable to improve upon a basic model for semantic relation identification, despite being pretrained on a large set of compounds. Their mediocre performance on the semantic relation classification task is likely caused by the use of individual word representations as the exclusive source of input, combined with the expectation that mathematical composition functions can directly extract and model patterns of interaction between pairs of words. We hypothesize that composition models can be improved by first modeling the semantic relations between words and then using the semantic relation representations together with the word representations as inputs to the composition process.

Acknowledgments

The author is indebted to Melanie Bell for the fruitful discussions and her comprehensive comments on the initial draft of the paper. The author would also like to thank Emanuel Dima and Erhard Hinrichs, as well as the anonymous reviewers for their insightful comments and suggestions. Financial support for the research reported in this paper was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center "The Construction of Meaning" (SFB 833), project A3.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1183–1193, Massachussetts, USA.

Laurie Bauer. 1983. *English word-formation*. Cambridge University Press.

Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011a. Torch7: A Matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011b. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Corina Dima and Erhard Hinrichs. 2015. Automatic noun compound interpretation using deep neural networks and word embeddings. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, pages 173–183, London, UK.

Georgiana Dinu, The Pham Nghia, and Marco Baroni. 2013a. DISSECT - DIStributional SEmantics Composition Toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 31–36, Sofia, Bulgaria.

Georgiana Dinu, The Pham Nghia, and Marco Baroni. 2013b. General estimation and evaluation of compositional distributional semantic models. In *ACL Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria.

David R. Dowty, Robert Wall, and Stanley Peters. 1981. *Introduction to Montague semantics*, volume 11 of *Synthese Language Library*. Springer Science & Business Media.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Mark Lauer. 1995. *Designing statistical language learners: Experiments on compound nouns*. Ph.D. thesis, Macquarie University.

Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan, and Israel Ramat-Gan. 2015. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*, Denver, CO, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015)*, Denver, CO, USA.

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(03):291–330.

Diarmuid Ó Séaghdha and Ann Copestake. 2013. Interpreting compound nouns with kernel methods. *Natural Language Engineering*, 19(03):331–356.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge. Published as University of Cambridge Computer Laboratory Technical Report 735.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Challenges in the Management of Large Corpora (CMLC-3)*.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 151–161. Association for Computational Linguistics.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, Uppsala, Sweden.

Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. Ph.D. thesis, University of Southern California.

Wenpeng Yin and Hinrich Schütze. 2014. An exploration of embeddings for generalized phrases. In *ACL 2014 Student Research Workshop*, pages 41–47, Baltimore, USA.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating Linear Models for Compositional Distributional Semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1263–1271.

A Overview of the Semantic Relations in the Tratz (2011) and Ó Séaghdha (2008) Datasets

Category name	Dataset percentage	Example
Objective		
OBJECTIVE	17.1%	leaf blower
Doer-Cause-Means		
SUBJECT	3.5%	police abuse
CREATOR-PROVIDER-CAUSE_OF	1.5%	ad revenue
JUSTIFICATION	0.3%	murder arrest
MEANS	1.5%	faith healer
Purpose/Activity Group		
PERFORM&ENGAGE_IN	11.5%	cooking pot
CREATE-PROVIDE-GENERATE-SELL	4.8%	nicotine patch
OBTAIN&ACCESS&SEEK	0.9%	shrimp boat
MITIGATE&OPPOSE	0.8%	flak jacket
ORGANIZE&SUPERVISE&AUTHORITY	1.6%	ethics authority
PURPOSE	1.9%	chicken spit
Ownership, Experience, Employment, Use		
OWNER-USER	2.1%	family estate
EXPERIENCER-OF-EXPERIENCE	0.5%	family greed
EMPLOYER	2.3%	team doctor
USER_RECIPIENT	1.0%	voter pamphlet
Temporal Group		
TIME-OF1	2.2%	night work
TIME-OF2	0.5%	birth date
Location and Whole+Part/Member of		
LOCATION	5.2%	hillside home
WHOLE+PART_OR_MEMBER_OF	1.7%	robot arm
Composition and Containment Group		
CONTAIN	1.2%	shoe box
SUBSTANCE-MATERIAL-INGREDIENT	2.6%	plastic bag
PART&MEMBER_OF_COLLECTION&CONFIG&SERIES	1.8%	truck convoy
VARIETY&GENUS_OF	0.1%	plant species
AMOUNT-OF	0.9%	traffic volume
Topic Group		
TOPIC	7.0%	travel story
TOPIC_OF_COGNITION&EMOTION	0.3%	auto fanatic
TOPIC_OF_EXPERT	0.7%	policy expert
Other Complements Group		
RELATIONAL-NOUN-COMPLEMENT	5.6%	eye shape
WHOLE+ATTRIBUTE&FEATURE &QUALITY_VALUE_IS_CHARACTERISTIC_OF	0.3%	earth tone
Attributive and Equative		
EQUATIVE	5.4%	fighter plane
ADJ-LIKE_NOUN	1.3%	core activity
PARTIAL_ATTRIBUTE_TRANSFER	0.3%	skeleton crew
MEASURE	4.2%	hour meeting
Other		
LEXICALIZED	0.8%	pig iron
OTHER	5.4%	contact lense
Personal*		
PERSONAL_NAME	0.5%	Ronald Reagan
PERSONAL_TITLE	0.5%	Gen. Eisenhower

Table 3: Semantic relations in the Tratz inventory - abbreviated version of Table 4.5 from Tratz (2011).

Relation	Frequency	Proportion	Examples
BE	191	13.2%	guide dog, rubber wheel, cat burglar
HAVE	199	13.8%	family firm, coma victim, sentence structure, computer clock, star cluster
IN	308	21.3%	pig pen, air disaster, evening edition, dawn attack
ACTOR	266	18.4%	army coup, project organiser
INST	236	16.4%	cereal cultivation, foot imprint
ABOUT	243	16.8%	history book, waterways museum, embryo research, house price

Table 4: Semantic relations in the Ó Séaghdha inventory - Table 6.2 from Ó Séaghdha (2008), augmented with examples from Table 3.1.